Evaluating the Performance of Climate Models Based on Wasserstein Distance

Vissio, G., Lembo, V., Lucarini, V. ORCID: https://orcid.org/0000-0001-9392-1471 and Ghil, M. (2020) Evaluating the Performance of Climate Models Based on Wasserstein Distance. Geophysical Research Letters, 47 (21). e2020GL089385. ISSN 0094-8276 doi: https://doi.org/10.1029/2020GL089385 Available at http://centaur.reading.ac.uk/93426/

Geophysical Research Letters

RESEARCH LETTER
10.1029/2020GL089385

Evaluating the Performance of Climate Models Based on Wasserstein Distance

Gabriele Vissio[1,2], Valerio Lembo[1,3], Valerio Lucarini[1,4,5], and Michael Ghil[6,7]

[1]CEN, Meteorological Institute, University of Hamburg, Hamburg, Germany, [2]Institute of Geosciences and Earth Resources (IGG), National Research Council (CNR), Turin, Italy, [3]Institute of Atmospheric Sciences and Climate (ISAC), National Research Council (CNR), Bologna, Italy, [4]Department of Mathematics and Statistics, University of Reading, Reading, UK, [5]Centre for the Mathematics of Planet Earth, University of Reading, Reading, UK, [6]Geosciences Department and Laboratoire de Météorologie Dynamique (CNRS and IPSL), Ecole Normale Supérieure and PSL University, Paris, France, [7]Department of Atmospheric & Oceanic Sciences, University of California, Los Angeles, CA, USA

Abstract We propose a methodology for intercomparing climate models and evaluating their performance against benchmarks based on the use of the Wasserstein distance (WD). This distance provides a rigorous way to measure quantitatively the difference between two probability distributions. The proposed approach is flexible and can be applied in any number of dimensions; it allows one to rank climate models taking into account all the moments of the distributions. By selecting the combination of climatic variables and the regions of interest, it is possible to highlight specific model deficiencies. The WD enables a comprehensive evaluation of the skill of a climate model. We apply this approach to a selected number of physical fields, ranking the models in terms of their performance in simulating them and pinpointing their weaknesses in the simulation of some of the selected physical fields in specific areas of the Earth.

Key Points:
- We propose a new method for evaluating the skill of a climate model
- The method allows for defining a ranking of climate models according to the selected variables of interest
- The method has the ability to highlight model deficiencies through emphasis on specific geographical regions and climatic variables

Supporting Information:
- Supporting Information S1

Correspondence to:
V. Lucarini,
v.lucarini@reading.ac.uk

Citation:
Vissio, G., Lembo, V., Lucarini, V., & Ghil, M. (2020). Evaluating the performance of climate models based on Wasserstein distance. *Geophysical Research Letters*, *47*, e2020GL089385. https://doi.org/10.1029/2020GL089385

Received 22 JUN 2020
Accepted 9 OCT 2020
Accepted article online 17 OCT 2020

1. Introduction and Motivation

Advanced climate models differ in the choice of prognostic equations and in the methods for their numerical solution, in the number of processes that are parametrized and the choice of the physical parametrizations, as well as in the way the models are initialized, to mention just their most important aspects. Comparing the performance of such models is still a major challenge for the climate modeling community (Held, 2005).

Model inadequacies, which may lead to large uncertainties in the model's predictions, result from structural errors—certain processes are incorrectly represented or not represented at all—as well as from parametric uncertainties, that is, the use of incorrect values for the parameters associated with processes that are correctly formulated in the model (Ghil & Lucarini, 2020; Lucarini, 2013). Intercomparing climate models and auditing them individually are essential for understanding which ones are more skillful in answering the specific climate question under study.

The need for testing systematically model performance has led the community to join forces through the Coupled Model Intercomparison Project (CMIP), which is currently in its sixth phase (Eyring et al., 2016). Dozens of modeling groups have agreed on a concerted effort to provide numerical simulations with standardized experimental protocols representative of specified climate forcing scenarios.

There is no standard suite of metrics to evaluate climate model performance nor, a fortiori, to decide whether a model does have skill in predicting future climate change. Lucarini et al. (2007) suggested that testing a climate model's performance requires considering a mixture of global and process-oriented metrics. Gleckler et al. (2008) proposed a multidimensional metric based on the comparison of the spatiotemporal variability of many climatic fields with respect to reference data sets and found that creating a scalar comprehensive metric is nontrivial. Eyring et al. (2016, 2020) have combined metrics and diagnostic tools designed to assess specific features of the climate system, while Lembo et al. (2019) have provided a tool to test the models' skill in representing the thermodynamics of the climate system.

Hence, it seems highly desirable to have a scalar metric that summarizes the information associated with model performance and that satisfies the mathematical axioms associated with the concept, as for the usual Euclidean distance. These axioms are listed in Text S1 of the supporting information (SI), and they are

satisfied by the root-mean-square distance, known as an L_2 metric in mathematics. The latter distance, though, is not appropriate for describing fully the difference between two distribution functions, while other metrics used in the climate sciences are not genuine distances, that is, they do not satisfy the axioms above.

We propose a new metric to assess a climate model's skill by taking into account every moment of a distribution and measuring the gap between it and another distribution of reference. The two distributions will be chosen here to describe model features, on the one hand, and the "real world", on the other, with the latter distribution being based on raw observations and/or a reanalysis thereof.

Ghil (2015) originally proposed the idea of using the Wasserstein distance (hereafter WD) (Dobrushin, 1970; Kantorovich, 2006; Villani, 2009) in the context of the climate sciences as a way to generalize the traditional concept of equilibrium climate sensitivity (Ghil & Lucarini, 2020) in the presence of a time-dependent forcing, such as seasonal or anthropogenic forcing. Robin et al. (2017) used the WD to compute the difference between the snapshot attractors of the Lorenz (1984) model for different time-dependent forcings, providing a link between nonautonomous dynamical systems theory and optimal transport. Vissio and Lucarini (2018) used the WD to evaluate the skill of a stochastic parametrization for a fast-slow system. Ning et al. (2014) proposed the use of the WD to quantify model error in variational data assimilation and presented an insightful application in the case of advection-diffusion dynamics with systematic errors in the velocity and diffusivity parameters. Please see Text S1 in the SI for further background on the WD.

Well-known WD drawbacks are (a) its computational requirements, which increase dramatically with the number of points used to construct the empirical distributions, and (b) the curse of dimensionality: The amount of data needed to explore accurately a higher dimensional phase space grows exponentially with the number of dimensions. Concerning (a), Vissio and Lucarini (2018) and Vissio (2018) have shown that the computational requirements are greatly reduced through data binning on a grid. As for (b), the WD will be calculated in a reduced phase space defined by the physical variables we wish to take into account in the evaluation of the model. The possibility of freely choosing the variables of interest makes the WD a flexible candidate for evaluating a climate model's skill.

The WD-based metric can complement the existing methods used for intercomparing climate models, such as ranking of model performances with respect to the root-mean-square error of the median of an ensemble (Flato et al., 2013) or weighted ensemble averaging schemes based on models' discrepancy from observations (Knutti et al., 2017). This letter is structured as follows. Data are presented in section 2, methods in section 3, results in section 4, and conclusions in section 5. The SI provides technical details.

2. Data

The WD methodology is presented in section 3. It is applied here to three climate fields:

- Near-surface air temperature;
- Precipitation; and
- Sea ice cover, computed from the sea ice area fraction.

The corresponding daily mean fields are available in the CMIP5 simulations for historical and RCP85 forcings (Taylor et al., 2012), and they are ranked with respect to the distance from reference daily data sets, specifically the European Centre for Medium-Range Weather Forecasts Re-Analysis (ERA) Interim for the temperature (Dee et al., 2011), Global Precipitation Climatology Project (GPCP) for the precipitation (Adler et al., 2003), and Ocean and Sea Ice - Satellite Application Facility (OSI-SAF) for the sea ice cover (EUMETSAT, Ocean and Sea Ice Satellite Application Facility, 2017). In order to further support the comparison and provide a benchmark, we analyzed the WD with respect to the National Center of Environmental Prediction (NCEP) Reanalysis 2 (Kanamitsu et al., 2002).

The fields are averaged on four distinct domains: (i) global, (ii) region between 30°S and 30°N (Tropics), (iii) region between 30°N and 90°N (northern extratropics), and (iv) Arctic—used only for sea ice extent. While temperature and precipitation analyses involve 30 models, taking into account sea ice extent allows to analyze just 22 models, due data availability.data sets. The time range spans 18 years, from 1 January 1997 to 31 December 2014. After the spatial averaging, the model data sets are obtained by concatenating the historical

runs, from 1997 to 2005, and the RCP85 runs, from 2006 to 2014. The acronyms of the models considered here are given in Table S1 of Text S2 in the SI.

The samples used in the WD calculations are drawn by performing a Ulam (1964) discretization of the phase space involved in each separate test. To do so, a regular grid is superposed over all the data sets used in the test, and its upper and lower limits, respectively, are fixed slightly above and below the maximum and minimum values among all the data sets used in it. Each dimension of the grid is then equally divided into 20 intervals; this yields 20^m m-dimensional cubes, where m is the number of fields taken into account in the test. These 20^m hypercubes provide the sample for each test. The results we present here are weakly sensitive to the specifics of the gridding. Nonetheless, a too coarse gridding removes a lot of the information we want to retain and analyze; a too fine gridding, instead, increases substantially the computing requirements, without making much statistical sense.

In order to highlight the flexibility and reliability of the method, we are going to calculate the WD distances in one-, two- and three-dimensional phase space and work with different field combinations averaged over distinct areas of the Earth.

3. Wasserstein Distance

Our objective is to create a ranking of the CMIP5 models based on their skill to reproduce the statistical properties of selected physical quantities. The reference distribution for these quantities is given by reanalysis and observational data sets, as explained in section 2; their WD to these data sets is a measure of the models' ability to reproduce these reference distributions. One can also describe this distance as the minimum "effort" to morph one distribution into the other (Monge, 1781). We present below a very simplified account of the theory.

The optimal transport cost (Villani, 2009) is defined as the minimum cost to move the set of n points from one distribution to another into an m-dimensional phase space. In the case of two discrete distributions, we write their measures μ and ν as

$$\mu = \sum_{i=1}^{n} \mu_i \delta_{x_i}, \qquad \nu = \sum_{i=1}^{n} \nu_i \delta_{y_i}; \tag{1}$$

here δ_{x_i} and δ_{y_i} are Dirac measures associated with a pair of points (x_i, y_i), whose fractional mass is (μ_i, ν_i), respectively, and $\sum_{i=1}^{n} \mu_i = \sum_{j=1}^{n} \nu_j = 1$, where all the terms in the sum are nonnegative. Using the definition of Euclidean distance

$$d(\mu, \nu) = \left[\sum_{i=1}^{n} (x_i - y_i)^2 \right]^{\frac{1}{2}}, \tag{2}$$

we can write down the quadratic WD for discrete distributions:

$$W_2(\mu, \nu) = \left\{ \inf_{\gamma_{ij}} \sum_{i,j} \gamma_{ij} [d(x_i, y_j)]^2 \right\}^{\frac{1}{2}}. \tag{3}$$

Here γ_{ij}, is a *transport protocol*, which defines how the fraction of mass is transported from x_i to y_j, while $d(x_i, y_j)$ is the Euclidean distance between a single pair of locations. The transport protocol realizing the minimum in Equation 3 is called the *optimal coupling*; see a visual explanation in Figure S1 of the SI.

We perform the Ulam discretization described above that allows us to shift from the distance between different distributions of points given by the time series to the distance between measures that can be estimated from such distributions (via data binning), while sticking to a discrete optimization problem, as discussed below. See Santambrogio (2015) for a survey of numerical methods for computing the WD. We thus proceed to quantify to what extent the measure of the observations and reanalysis from section 2, projected on the variables of interest, differs from the corresponding measures for the climate models.

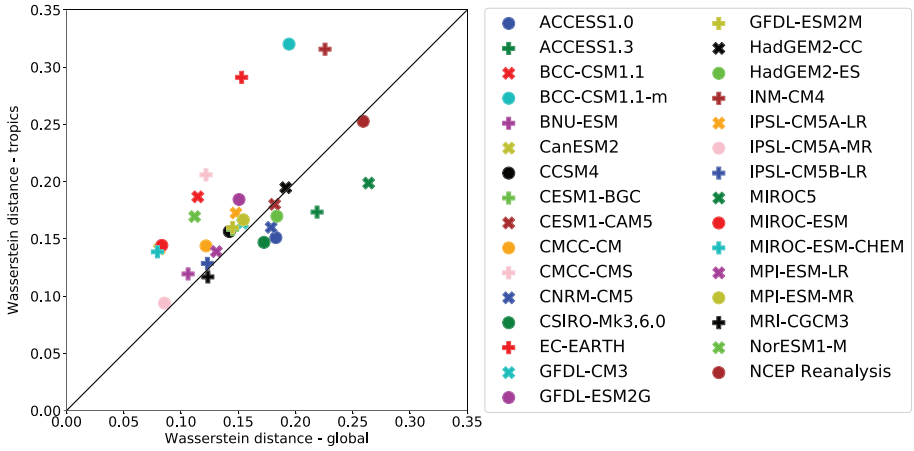

Figure 1. Two-dimensional WD for the temperature and precipitation fields, averaged over the globe (horizontal axis) and over the Tropics (vertical axis). The acronyms of the models used are spelled out in Text S2 of the SI.

The estimate of the coarse-grained probability of being in a specific grid box is given by the time fraction spent in that box (Ott, 1993; Strogatz, 2015). In fact, the WD does provide robust results even with a very coarse grid (Vissio, 2018; Vissio & Lucarini, 2018). Therefore, in the case at hand, the locations x_i and y_j will indicate the cubes' centroids, while γ_{ij} indicate the corresponding densities of points. If $(k_1, k_2, \dots, k_m)$ are indices running from 1 to n, the cube with position x in the m-dimensional space will be identified by the m-tuple $k_1^x, k_2^x, \dots, k_m^x$. We then define $d(x,y)^2 = \sum_{l=1}^{m}(k_l^x - k_l^y)^2$. To further simplify the computations, we exclude all the grid boxes containing no points at all. Finally, we divide the distance by n; therefore, the one-, two-, and three-dimensional WDs take values between a minimum of 0 and a maximum equal to 1, $\sqrt{2}$, and $\sqrt{3}$, respectively.

We used a suitably modified version of the Matlab software written by G. Peyré—available at https://nbviewer.jupyter.org/github/gpeyre/numerical-tours/blob/master/matlab/optimaltransp_1_linprog.ipynb—to perform the calculations. The modifications include the data binning and the estimation of the measures, as well as adapting to a dimension $m \geq 2$.

4. Ranking the Models

Figure 1 shows the WD calculated in the two-dimensional phase space composed by the temperature and precipitation fields, averaged over the whole Earth and the Tropics, for each CMIP5 model. In order to provide a benchmark, we chose to include the WD results between the NCEP reanalysis and the references given by the ERA temperature and GPCP precipitation fields, respectively.

Somewhat surprisingly, the NCEP reanalysis yields the largest values in both distances. Thus, the average CMIP5 distance to the combined ERA-and-GPCP reference data sets is 0.149, while the NCEP distance is 0.259, exceeded only by the value 0.264 given by the MIROC5 model; see Table S1 in the SI for the list of models. Note that the one-dimensional WDs of the NCEP reanalysis for the globally averaged temperature and precipitation equal 0.033 and 0.255, respectively, which indicates the inadequacy of the NCEP data set in representing the statistics of precipitation. Despite the well-known difficulties with simulating the very rough precipitation field by using the still fairly coarse CMIP5 models (Mehran et al., 2014; Neelin et al., 2013), the results point to the overall accuracy reached by CMIP5 simulations when dealing with global averages of temperatures and precipitation.

We evaluate next the problems still encountered by CMIP5 models in reproducing key aspects of tropical dynamics (Tian & Dong, 2020). Averaging the data over the Tropics, we obtain the ranking on the vertical axis in Figure 1. The WD distance is for most data sets larger than when looking at globally averaged quantities (the models' mean is 0.173) and underline the poorer CMIP5 model performances in this region. With few exceptions, the models seem less reliable in the Tropics, where three of the models exceed the NCEP reanalysis distance. This distance is very similar to what has been found for the globally averaged case.

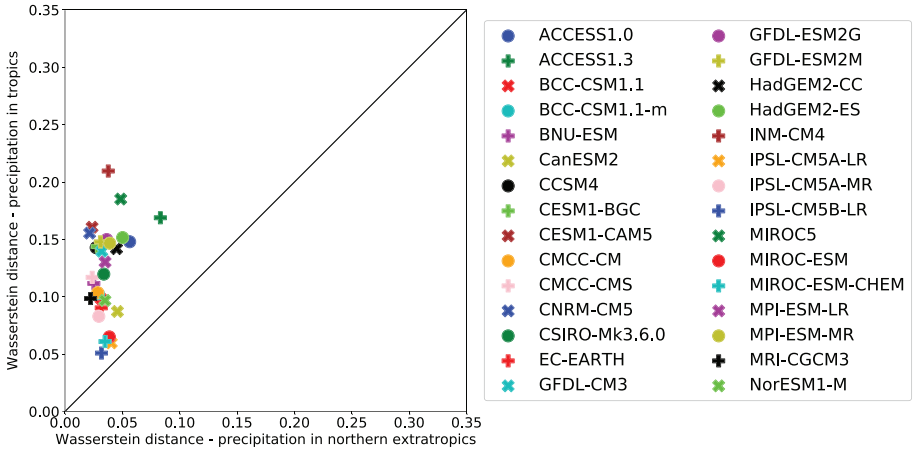

Figure 2. One-dimensional WD for precipitation averaged over the northern extratropics (from 30°N to 90°N) on the horizontal axis and over the Tropics (from 30°S to 30°N on the vertical axis).

Next, we show how the WD can be used to perform comparative analyses of the performance of a given model or of a group of models with respect to different climatic observables. Focusing on the relative performance of temperature and precipitation in the Tropics versus the Northern Hemisphere extratropics, Figures 2 and 3 illustrate one-dimensional WDs computed in the former versus the latter region. Using the diagonal line indicating equal values for the two distances as a reference, we can easily check in Figure 2 that, for all CMIP5 models, the precipitation field is less well reproduced in the Tropics than in the extratropics: It is extremely challenging to reproduce accurately the statistics of by-and-large convection-driven precipitation, since the choice of the parametrization schemes and their tuning plays an essential role. The situation for the temperature field is similar but less uniformly so while in Figure 2 all the results cluster above the diagonal but roughly below WD $\simeq$0.2, the scatter in Figure 3 is larger, with some results below the diagonal and some between $0.2 \lesssim \mathrm{WD} \lesssim 0.3$.

Figure 4 shows the scatter diagram of one-dimensional WDs for the precipitation in the Tropics versus the WDs of sea ice extent in the Arctic. Arctic sea ice cover is a very important indicator of the state of both hydrosphere and cryosphere, as well as of their mutual coupling; it is overestimated in CMIP5 models during the winter and spring seasons (Flato et al., 2013; Randall et al., 2007).

Figure 4 demonstrates that the sea ice cover in the models is closer to the observations than the tropical precipitation in 12 CMIP5 models out of the 22 examined. Nevertheless, seven models better describe tropical

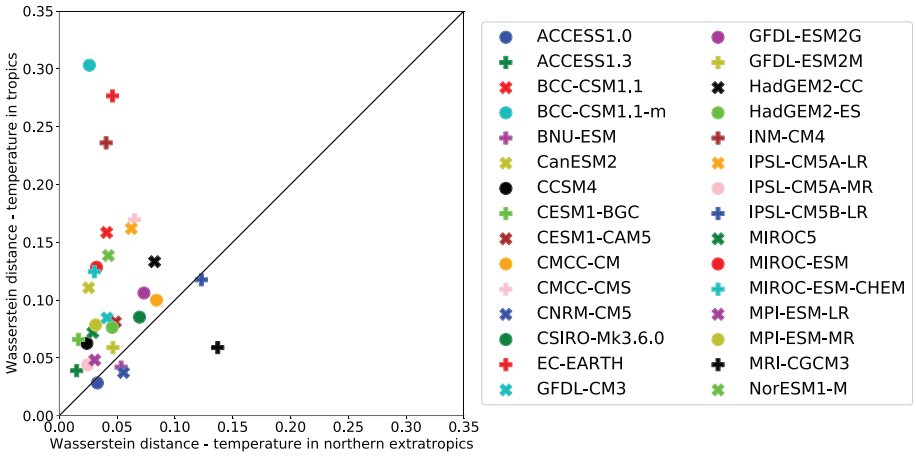

Figure 3. Same as Figure 2 but for the temperature field.

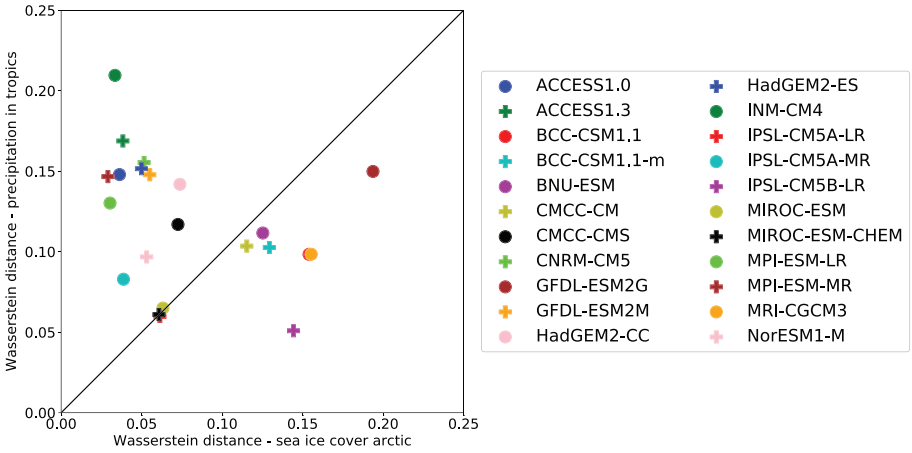

Figure 4. One-dimensional WDs of average precipitation in the Tropics versus the average sea ice extent in the Arctic.

precipitation than sea ice extent in the Arctic, while three models have a similar—and relatively low—WD for both fields. This test indicates that a correct representation of the statistics of these two fields is quite challenging for the CMIP5 models.

We compare next the performance of the CMIP5 models with respect to three different rankings. First, the three-dimensional WD is computed taking into account three physical quantities: globally averaged temperature and precipitation, along with sea ice extent in the Arctic. Note that, to ease the interpretation of Figure 5, the models are listed on the vertical axis according to the rank provided by this methodology.

The model ranking introduced herein is further compared with the rankings based on the first two moments of the distribution of reference. For each of the three physical quantities above, we compute the normalized mean, taking the absolute value of the difference between the mean of the distribution of the model field and that of the reference field and dividing this difference by the standard deviation of the distribution of reference. The three means for the three fields are then averaged, and the same procedure is repeated for the normalized standard deviation.

We can see that the models' performance is quite different depending on the ranking being used. As an example, we focus on the BCC-CSM1.1 and BCC-CSM1.1-m models. The ranking based on the mean shows a rather good performance for both, with positions 7 and 10, respectively; nevertheless, they occupy positions 16 and 21 in the WD ranking. The latter low positions are due to their bad performances when it comes to standard deviation, where the two come last.

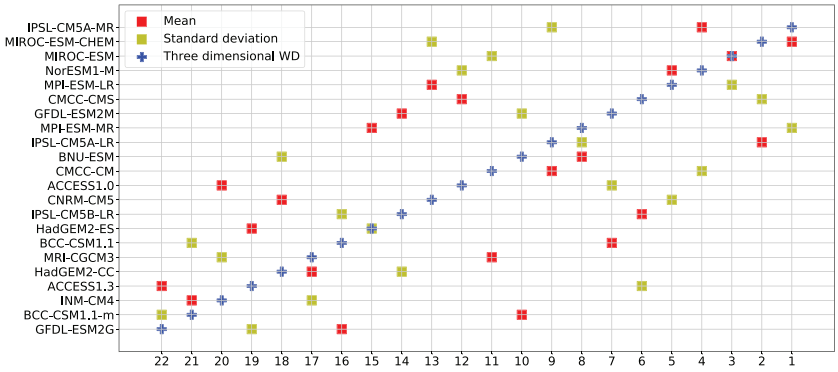

Figure 5. Comparing 22 CMIP5 models (vertical axis) versus their positions in the ranking (horizontal axis): three-dimensional WD—heavy blue "+" sign, mean—red filled square, and standard deviation—yellow filled square. See text for explanations. See Tables S2–S4 in the SI for detailed results.

The reverse instance is also clear by looking at those models that, while performing well in terms of variability, occupy lower rankings based on the WD due to their poor performance in the mean; see, for instance, the case of MPI-ESM-MR, with Position 1 in the standard deviation, 8 in WD, and 15 in the mean. The WD score accounts for the information carried by the whole distribution—that is, by the mean, standard deviation and higher moments—and clearly balances out the first and second moments thereof.

A more peculiar instance is provided by HadGEM2-CC and HadGEM2-ES, which rank in this order for both the mean (17th and 19th) and the standard deviation (14th and 15th) but in the reverse order in the WD ranking (18th and 15th). This apparent paradox could be due to the presence of nontrivial second-order correlations between the variables or from the effect of higher moments of the distributions.

Note that, for the 18 year time interval studied herein (1997–2014), the results obtained applying the WD approach in three-dimensional phase space are not very different from those given by averaging the three corresponding one-dimensional distances. This agreement is due to the unimodality of the distributions taken into account, and things would be different in the case of multimodal distributions. In any case, the full application of the multidimensional WD leads to more robust results, as all correlations between the variables are taken into consideration.

5. Conclusions

We have proposed a new methodology to study the performance of climate models based on the computation of the WD between the multidimensional distributions of suitably chosen climatic fields of reference data sets and those of the models of interest. This method takes into account all the moments of the distributions, unlike most evaluation methods for climate models used so far, which consider solely the first two moments of the distribution. It is, therefore, more informative and takes into account also the distribution of extreme events. The methodology allows one to consider several variables at the same time, and it helps select such variables depending on the goal of the intercomparison. Thus, it can assist in disentangling the correlation between different climatic quantities.

The proposed methodology has been proven to be effective in pointing to climate modeling problems related to the representation of quantities like precipitation or sea ice extent over limited areas, such as the Tropics and the Arctic, respectively; see again Figures 2 and 3. Furthermore, this methodology can be applied to studying model performance for a given climatic variable over different spatial domains, as seen in Figures 1–4, as well as relative model performance for different fields, as seen in Figure 4. This flexibility can help guide attempts at model improvements by providing robust diagnostics of the least well simulated field—temperature, precipitation, or sea ice extent—or region, namely, hemisphere, the Tropics, or the Arctic.

Throughout the paper, we have shown the application of this approach to different physical fields, providing a ranking of CMIP5 models for specific sets of fields, as well as a way to highlight model weaknesses to help focus the honing of climate models. Getting more reliable models will lead to better simulations and, therefore, to more accurate climate predictions.

Acknowledgments

The authors thank the climate modeling groups for producing and making available their model output and acknowledge the World Climate Research Programme Working Group on Coupled Modelling. The authors thank E. Foufoula-Georgiou, T. T. Georgiou, and C. Kuehn for bringing to their attention relevant papers. V. Lu thanks N. Gigli and F. Santambrogio for inspiring exchanges. V. Lu and V. Le have been supported by the DFG-CRC TRR181 (Project 274762653). V. Lu and M. G. acknowledge the support received from the H2020 project TiPES (Grant 820970). The present paper is TiPES contribution #30. Work on this paper has also been supported by the EIT Climate-KIC (Grant 190733).

Data Availability Statement

GPCP and NCEP reanalysis 2 data have been provided by NOAA/OAR/ESRL PSD (https://www.esrl.noaa.gov/psd/). The EUMETSAT OSI-SAF 1979-2015 sea ice concentration (v2.0, 2017) has been provided by ICDC, University of Hamburg (https://icdc.cen.uni-hamburg.de/en/seaiceconcentration-osisaf.html). CMIP5 data sets are accessible online (https://esgf-data.dkrz.de). The original data obtained in this paper are available online (https://figshare.com/articles/dataset/VissioLemboLucariniGhil2020_zip/12982406).

References

Adler, R. F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P., Janowiak, J., et al. (2003). The version 2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979–present). *Journal of Hydrometeorology*, *4*, 1147–1167. https://doi.org/10.1175/1525-7541(2003)004<1147:TVGPCP>2.0.CO;2

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

Dobrushin, R. L. (1970). Prescribing a system of random variables by conditional distributions. *Theory of Probability & Its Applications*, *15*(3), 458–486.
EUMETSAT, Ocean and Sea Ice Satellite Application Facility (2017). Global sea ice concentration climate data record 1979-2015 (v2.0, 2017). Norwegian and Danish Meteorological Institutes. http://osisaf.met.no
Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., et al. (2020). Earth system model evaluation tool (esmvaltool) v2.0—An extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of earth system models in cmip. *Geoscientific Model Development*, *13*(7), 3383–3438. https://doi.org/10.5194/gmd-13-3383-2020
Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*, 10,539–10,583. https://doi.org/10.5194/gmd-9-1937-2016
Eyring, V., Righi, M., Lauer, A., Evaldsson, M., Wenzel, S., Jones, C., et al. (2016). ESMValTool (v1.0)—A community diagnostic and performance metrics tool for routine evaluation of Earth system models in CMIP. *Geoscientific Model Development*, *9*, 1747–1802. https://doi.org/10.5194/gmd-9-1747-2016
Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., et al. (2013). Evaluation of climate models. In T. F. Stocker et al. (Eds.), *Climate change 2013: The physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change*. Cambridge, UK and NY, USA: Cambridge University Press.
Ghil, M. (2015). A mathematical theory of climate sensitivity or, How to deal with both anthropogenic forcing and natural variability? In C.-P. Chang, M. Ghil, M. Latif, J. M. Wallace (Eds.), *Climate change: Multidecadal and beyond* (Vol. 6, pp. 31–52). Singapore: World Scientific Publishing Co.
Ghil, M., & Lucarini, V. (2020). The physics of climate variability and climate change. *Reviews of Modern Physics*, *92*, 035002. https://doi.org/10.1103/RevModPhys.92.035002
Gibbs, A. L., & Su, F. E. (2002). On choosing and bounding probability metrics. *International Statistical Review*, *70*(3), 419–435. https://doi.org/10.1111/j.1751-5823.2002.tb00178.x
Gleckler, P. J., Taylor, K. E., & Doutriaux, C. (2008). Performance metrics for climate models. *Journal of Geophysical Research*, *113*, D06104. https://doi.org/10.1029/2007JD008972
Halmos, P. R. (2017). *Introduction to Hilbert Space and the Theory of Spectral Multiplicity*. Courier Dover Publications.
Held, I. M. (2005). The gap between simulation and understanding in climate modeling. *Bulletin of the American Meteorological Society*, *86*, 1609–1614. https://doi.org/10.1175/BAMS-86-11-1609
Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S.-K., Hnilo, J. J., Fiorino, M., & Potter, G. L. (2002). NCEP-DOE AMIP-II reanalysis (R-2). *Bulletin of the American Meteorological Society*, *83*, 1631–1643. https://doi.org/10.1175/BAMS-83-11-1631
Kantorovich, L. V. (2006). On the translocation of masses. *Journal of Mathematical Sciences*, *133*(4), 1381–1382. originally published in Doklady Akademii Nauk SSSR, 37 (78), 199–201 (1942).
Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., & Eyring, V. (2017). A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters*, *44*, 1909–1918. https://doi.org/10.1002/2016GL072012
Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.
Lembo, V., Lunkeit, F., & Lucarini, V. (2019). Thediato (v1.0)—A new diagnostic tool for water, energy and entropy budgets in climate models. *Geoscientific Model Development*, *12*(8), 3805–3834. https://doi.org/10.5194/gmd-12-3805-2019
Lorenz, E. N. (1984). Irregularity: A fundamental property of the atmosphere. *Tellus A*, *36*(2), 98–110.
Lucarini, V. (2013). Modeling complexity: The case of climate science. In U. Ghde, S. Hartmann, J. H. Wolf (Eds.), *Models, simulations, and the reduction of complexity* (pp. 229–254). Hamburg: De Gruyter.
Lucarini, V., Calmanti, S., Dell'Aquila, A., Ruti, P. M., & Speranza, A. (2007). Intercomparison of the Northern Hemisphere winter mid-latitude atmospheric variability of the IPCC models. *Climate Dynamics*, *28*(7), 829–848. https://doi.org/10.1007/s00382-006-0213-x
Mehran, A., AghaKouchak, A., & Phillips, T. J. (2014). Evaluation of CMIP5 continental precipitation simulations relative to satellite-based gauge-adjusted observations. *Journal of Geophysical Research: Atmospheres*, *119*, 1695–1707. https://doi.org/10.1002/2013JD021152
Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. Histoire de l'Acadmie Royale des Sciences, 666–704.
Neelin, J. D., Langenbrunner, B., Meyerson, J. E., Hall, A., & Berg, N. (2013). California winter precipitation change under global warming in the coupled model intercomparison project phase 5 ensemble. *Journal of Climate*, *26*(17), 6238–6256. https://doi.org/10.1175/JCLI-D-12-00514.1
Ning, L., Carli, F. P., Ebtehaj, A. M., Foufoula-Georgiou, E., & Georgiou, T. T. (2014). Coping with model error in variational data assimilation using optimal mass transport. *Water Resources Research*, *50*, 5817–5830. https://doi.org/10.1002/2013WR014966
Ott, E. (1993). *Chaos in dynamical systems*. Cambridge, UK: Cambridge University Press.
Randall, D. A., Wood, R. A., Bony, S., Colman, R., Fichefet, T., Fyfe, J., et al. (2007). Climate models and their evaluation. In S. Solomon et al. (Eds.), *Climate change 2007: The physical science basis. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change*. Cambridge, UK and NY, USA: Cambridge University Press.
Robin, Y., Yiou, P., & Naveau, P. (2017). Detecting changes in forced climate attractors with Wasserstein distance. *Nonlinear Processes in Geophysics*, *24*, 393–405. https://doi.org/10.5194/npg-24-393-2017
Santambrogio, F. (2015). *Optimal transport for applied mathematicians: Calculus of variations, PDES, and modeling*, *Progress in Nonlinear Differential Equations and Their Applications*. Basel: Birkhäuser.
Strogatz, S. H. (2015). *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering, 2nd edition*. Boulder, CO: Westview Press.
Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*, 485–498. https://doi.org/10.1175/BAMS-D-11-00094.1
Tian, B., & Dong, X. (2020). The double-ITCZ bias in CMIP3, CMIP5, and CMIP6 models based on annual mean precipitation. *Geophysical Research Letters*, *47*, e2020GL087232. https://doi.org/10.1029/2020GL087232
Ulam, S. M. (1964). *Problems in modern mathematics, A collection of mathematical problems*. New York: Science Edition Wiley.
Villani, C. (2009). *Optimal transport: Old and new*. Berlin Heidelberg, Germany: Springer-Verlag.
Vissio, G. (2018). *Statistical mechanical methods for parametrization in geophysical fluid dynamics*, *Reports on Earth System Science* (Vol. 212). Hamburg: Max-Planck-Institut für Meteorologie.
Vissio, G., & Lucarini, V. (2018). Evaluating a stochastic parametrization for a fast-slow system using the Wasserstein distance. *Nonlinear Processes in Geophysics*, *25*, 413–427. https://doi.org/10.5194/npg-25-413-2018